DOCUMENT RESUME

ED 298 965 IR 052 506

AUTHOR Olson, Dennis H.; Richards, Mary
TITLE The Effectiveness of Keyword Searching in an Online Catalog To Access a Special Collection.
PUB DATE [86]
NOTE 17p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bibliographic Records; Comparative Analysis; Indexing; Information Retrieval; *Online Catalogs; *Relevance (Information Retrieval); *Search Strategies; *Subject Index Terms; Vocational Rehabilitation
IDENTIFIERS *Keywords; *MARC

ABSTRACT

This study was conducted to compare the effectiveness of keyword searching in an online catalog to traditional subject searching in one special collection of vocational rehabilitation materials. Subject area specialists assessed the contents of books using their terminology, and compared the results to MARC records. It was found that slightly under half of the items from the sample could be accessed using subject searching, while over 86% of the records could be retrieved when all MARC fields were searched by keyword; the title field provided the greatest access. The prevalence of multiword terms in the sample indicates that a catalog capable of proximity searching would be very useful. Two figures are included. (13 bibliographic notes/references) (Author/EW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED298965

THE EFFECTIVENESS OF KEYWORD SEARCHING IN AN ONLINE CATALOG TO ACCESS A SPECIAL COLLECTION

By Dennis H. Olson, Reference Librarian, University of Wisconsin-Stout, Menomonie, WI and Mary Richards, Head of Cataloging, University of Wisconsin-Stout, Menomonie, WI.

IR052506

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Dennis H. Olson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ABSTRACT

A study was done to compare the effectiveness of keyword searching in an online catalog to traditional subject searching in one special collection vocational rehabilitation. Subject area specialists assessed the contents of books using their terminology and compared results to MARC records. Slightly under half of the items from the sample could be accessed using subject searching while over 86% of the records could be retrieved when all MARC fields were searched by keyword. The title field provided the greatest access. The prevalence of multiword terms in the sample indicates that a catalog capable of proximity searching would be very useful.

THE EFFECTIVENESS OF KEYWORD SEARCHING IN AN ONLINE CATALOG TO ACCESS A SPECIAL COLLECTION

By Dennis Olson and Mary Richards

INTRODUCTION

Librarians who work with special subject collections often find that Library of Congress Subject Headings are inadequate as access points for these collections. The University of Wisconsin-Stout Library Learning Center contains special collections in several subject areas including vocational rehabilitation, manufacturing, hotel and restaurant management, tourism, and graphic communication. Access to the materials in each of these collections can be difficult. Materials on a subject such as vocational rehabilitation tend to be listed under a few broad subject headings in our public catalog while students and faculty describe their information needs using specific terms such as work evaluation, work adjustment or caseload management that are not represented by an LC Heading and thus not in our catalog.

This was demonstrated in a most striking way several years ago when a faculty member from the Vocational Rehabilitation Department brought his class in for library instruction. Each member of the class was asked to do a report on a different topic within the field of vocational rehabilitation. Much to our embarrassment most of the 25-30

topics in the assignment were represented only by the heading vocational rehabilitation in our catalog.

Our present online catalog is limited to searches by author, title, and subject but in a short time we will be converting to a catalog that offers keyword searching. We assume that keyword searching will be very popular once it is available based on the experiences of others. For example, librarians at Mankato State University reported that keyword searching accounted for nearly half of all of the searches performed on their online catalog.1 Our hypothesis is that keyword searching would also provide improved access to our special collections such as vocational rehabilitation, primarily because it would allow users to enter the terms common to their field. It is important to point out that when we speak of keyword searching we refer to the capacity of an online catalog to search the bibliographic records for the occurence of one or more terms chosen by the user and to post the results of this search.

LITERATURE REVIEW

The keyword index is considered to date back as far as the 16th century in western Europe.2 Computer technology was first applied to keyword indexing in 1958 to develop the keyword in context or KWIC index.3 Studies of keyword searching using computer technology began to appear in the literature in the late 1950's, with the investigations falling generally into two main categories:

The first category involved comparisons of automated keyword indexing against indexing done by professional indexers to measure whether the machine would come up with the same terms. The traditional human-assigned index was used as the standard against which to measure automated indexing. Neil Yerkey of the University of Akron examined several such studies and generalized that keyword indexes work "at least about 50% as well as conventional indexes."[4] Pauline Hodges in a study at Ohio State University also reported a 50% retrieval rate.[5]

The second major group involved studies of titles in particular academic disciplines to determine whether they had sufficient significant terms to be useful for keyword indexing. Results of these studies were not entirely consistent. Hodges said that social sciences literature had the best retrieval rate.[6] Others found better results in the sciences. Hilda Feinberg in a comparative study of title derivative indexing techniques, concluded that titles can be used to produce acceptable indexes for a broad spectrum of disciplines.[7]

In recent years, interest in keyword concepts has been renewed because of applications in on-line catalogs. Pauline Cochrane reported on two studies conducted by graduate students at Syracuse University that are relevant to our research.[8] In the first, Judith Hudson compared keyword searching to class number searching for geographic area materials searching her results indicated that title

and subject fields taken together provided adequate access, but neither was sufficient alone. In the second, Michael Ubaldini expanding on Hudson's work, compared the effectiveness of the various MARC fields for keyword searching. He ranked title, subject headings, notes and main entry as the most useful fields, in that order.

Another approach was taken in a recent study conducted at the University of Western Ontario by Alexis Jamieson, Elizabeth Dolan and Luc Declerck.9 They looked at the possibility of using keyword searching as an alternative to a built-in cross reference structure. They checked non-preferred headings in LC authorities against MARC records to see if the terms used as cross-references would be found in any MARC fields. They concluded that keyword searching was not a substitute for authority control.

Like these previous studies, our research was designed to compare the effectiveness of traditional subject searching and keyword searching. In addition, we have attempted to identify the parts of the bibliographic record that were most useful for keyword searching. We chose to limit our sample to a particular field of study that had inadequate subject access as determined by user complaints. We hoped that this information would help us to determine whether new catalog technology would improve our access to these materials or other whether we should pursue other methods to do this.

METHODOLOGY

A sample of 116 items was created by selecting every fifth item from the call number ranges that encompassed the majority of the vocational rehabilitation materials. The Marc records were printed for each item in the sample and each was given a unique item number.

Since the Library of Congress Subject Headings were known to be inadequate for this subject area, they could not be used as the standard against which keyword searching was measured. A new benchmark was created by hiring two graduate students in vocational rehabilitation to simulate the search strategies of a sophisticated user. They worked independently to examine each book in the sample and assign descriptors to them. They were asked to use the same terms that professionals in the field would use to describe the content of the items.

The students received training in basic indexing skills and practiced on items from outside the sample before beginning work on the project. A thesaurus of vocational rehabilitation terms would have been useful but the only thesaurus we could locate had over 1,000 terms, too many for our purposes.

The resulting data includes two separate lists of descriptors assigned by the indexers which could be compared against the search terms available in a keyword search of the Marc records. The data was analyzed to determine which fields contained the chosen terms, which fields provided the only access points for a given record, how many items were

inaccessible, and which terms assigned by the indexers would have been matched by see references.

RESULTS

The terms used by the indexers in our study matched Library of Congress subject headings for slightly under half (48.6%) of the records. Figure 1 shows how the keyword search capability increased accessibility of the records by the identified index terms. Searching the subject field by keyword provided only a modest improvement over traditional subject searching (53.7% versus 48.6%). The success rate increased significantly, to 70.3% of the records accessed by the chosen terms when the title field was searched.

The third circle illustrates that 86.1% of the records could be retrieved when all MARC fields were searched. These results demonstrate a substantially improved rate of access using a keyword approach, with the title field showing the greatest potential for improving subject access. In fact, for 19% of the records keyword access to the title provided the only access to the item using the terms selected by the indexers in our study.

Figure 2 shows in more detail the usefulness of the various fields in the MARC record for keyword searching. The graph indicates the percentage of records for each MARC field that contained the terms assigned by the indexers. Title (245) and subject heading (650) fields are demonstrated to be the richest for keyword searching by a wide margin, echoing the findings of the Syracuse study.

Fields containing corporate names (110, 260, 710) provided some matches. The remaining fields represented notes and alternate forms of titles such as series (440, 490) and titles traced differently (740). Although notes (505) and alternate title fields would seem to have high potential for significant terms, their relatively low representation in the statistics reflects their low numbers in the sample. Their representation is quite low in the general population of records as well. Crawford, Stovel, and Bales report that in the RLIN database, only 13.8% of the records in all formats (except archives and manuscripts) contained a 440 (series title) field and only 4.37% of records included a 505 (contents note) field.10 In our sample 25% of series title fields (440) that were present contained significant search terms along with 17% of the note fields (505).

PROBLEMS

It is important that we acknowledge the influence that titles may have on indexers when they assign index terms. It is very likely that the indexers in our study were influenced by the titles of the items when they assigned the index terms. This could account for some but certainly not all of the 70.3% of the records that were accessible by terms found in the titles. We must also acknowledge that although it would seem that keyword searching offers enhanced access to our vocational rehabilitation materials it is not without problems. As Rothman points out keyword

searching places a greater burden on the user to include all relevant keywords and variants in order to do a comprehensive search.11 The user must also provide the linkage between terms that is provided by the syndetic structure contained in the Library of Congress Subject Headings.

One finding that we had not anticipated was that of the 634 total terms assigned 603, or 95%, were in the form of phrases rather than words. Multi-word terms like work evaluation, sheltered workshops, rehabilitation counselors, and work adjustment were common. This poses a problem for anyone searching by keyword because of the inability of a simple keyword index to pull these multiword concepts together.12

This prevalence of phrases would suggest that to maximize the effectiveness of keyword searching, online systems require the capacity to do proximity searching. It is surprising then that of the 48 online catalogs examined by Joe Mathews only 7 offered proximity searching.13 Lacking this capability, the searcher would have to choose one of the words to initiate the search and then if the system allows, limit the search to records containing the second word as well. With many phrases, this narrowing technique would produce a very high rate of recall and a low rate of precision. For example, a search on a subject like work evaluation would produce many records that had both of these words in them but not in that sequence. Limiting the

search to those items that contained the words in the same field would be an improvement but still problematic. The most precise search strategy would allow us to search "work" next to "evaluation."

IMPLICATIONS

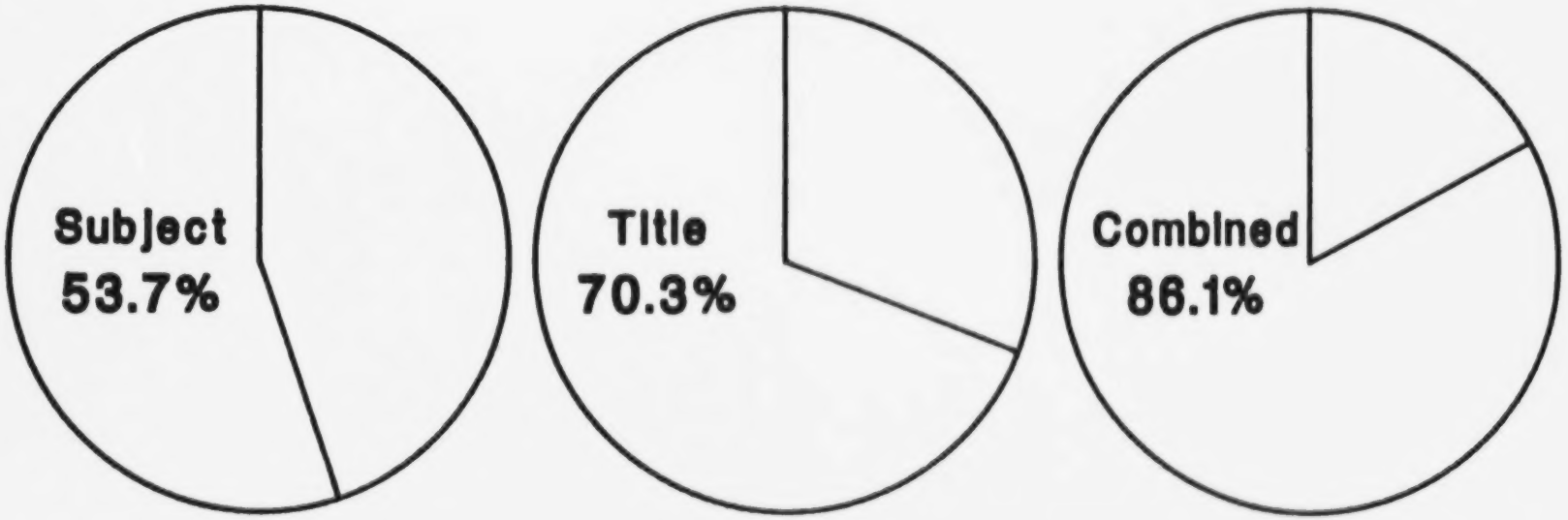
Our study indicates that keyword searching would improve the subject access to materials in our vocational rehabilitation collection. It would be interesting to know whether users who approach the catalog would experience the same level of success searching by keyword as the indexers in our study have indicated. The speed in which users could perform their searches and the rate of precision appears to depend on the capability of the system to offer proximity searching.

Further study should be done to see if the results are the same for other specialized subject areas. Online use studies measuring user success levels with keyword searching versus searching by Library of Congress Subject Headings would help librarians determine whether resources should be allocated to enhance records with keywords, improve search software, or compensate by making users aware of problems at the reference desk and through instruction.

REFERENCES

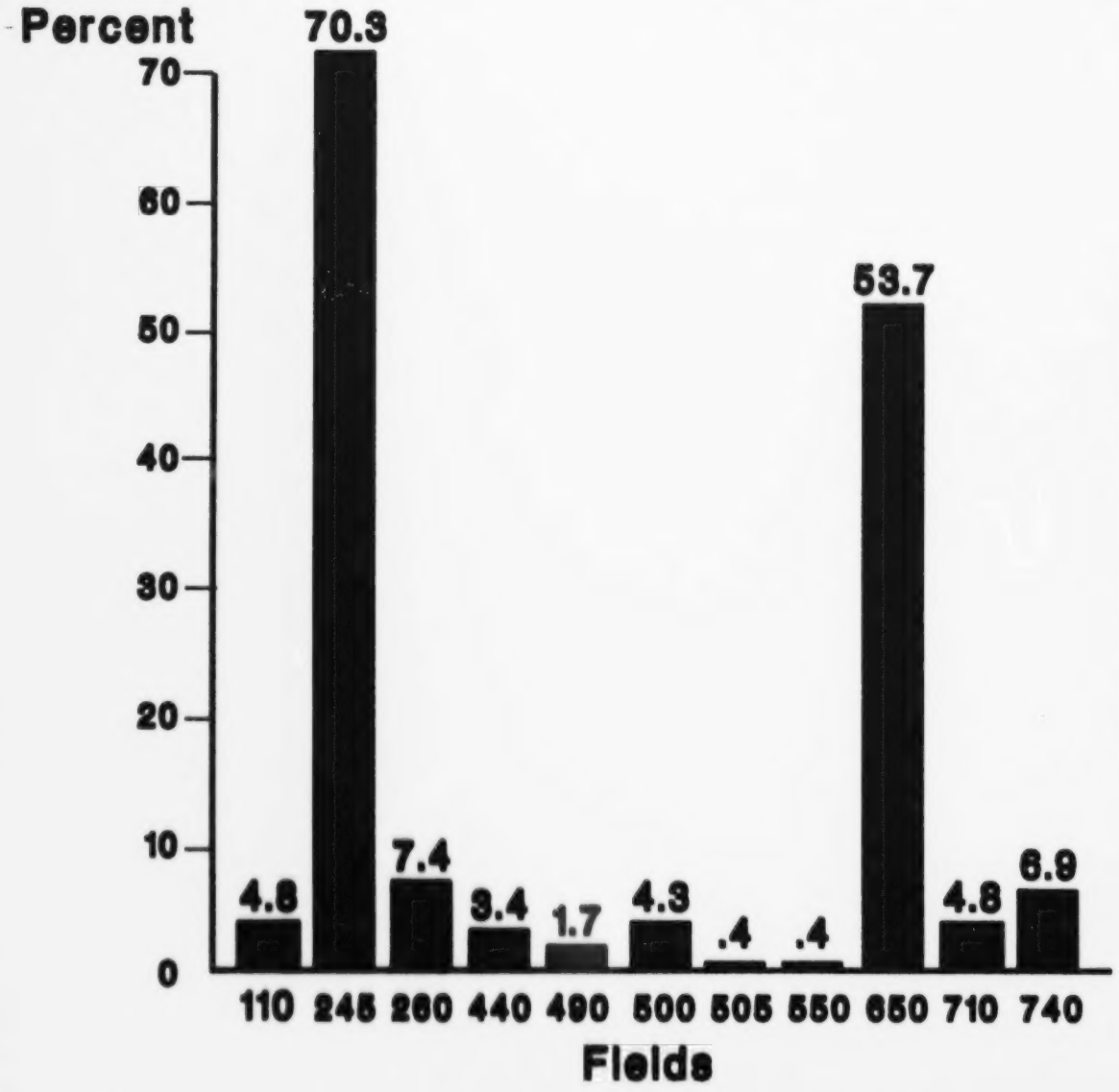
1Joseph R. Matthews, Gary S. Lawrence, and Douglas K. Ferguson, eds., Using Online Catalogs: A Nationwide Survey (New York: Neal-Schuman, 1983).
2John Rothman, "Is Indexing Obsolete? Keyword Indexing and Free-Text Searching," in Indexing Specialized Formats and Subjects, ed. Hilda Feinberg (Metuchen, N.J.: Scarecrow, 1983).
3Karen Humbert, "To be KWIC or not to be KWIC," Drexel Library Quarterly 8 (April 1972): 149-157.
4Neil Yerkey, "Models of Index Searching and Retrieval Effectiveness of Keyword-in-Context Indexes," Journal of the American Society for Information Science 24 (August 1973): 282-285.
5Pauline R. Hodges, "Keyword in Title Indexes: Effectiveness of Retrieval of Computer Searches," Special Libraries 74 (January 1983): 56-60.
6Ibid., p. 56.
7Hilda Feinberg, Title Derivative Indexing Techniques: A Comparative Study (Metuchen, N.J.: Scarecrow, 1973).
8Pauline Cochchrane, Redesign of Catalogs and Indexes for Improved Online Subject Access (Phoenix: Oryx, 1985).
9Alexis J. Jamieson, Elizabeth Dolan, and Luc Declerck, "Keyword Searching vs. Authority Control in an Online Catalog," Journal of Academic Librarianship 12 (November 1986): 277-283.
10Walt Crawford, Lennie Stovel, and Kathleen Bales, Bibliographic Displays in the Online Catalog (White Plains, N.Y.: Knowledge Industry Publication 1986).
11John Rothman, Indexing Specialized Formats and Subjects.
12Elaine P. Burress, "Automated Indexing Versus KWOC: A Performance Comparison," Journal of the American Society for Information Science 31 (January 1980): 60-61.
13Joseph R. Matthews, Using Online Catalogs.

Figure 1



Success rate of keyword searches in subject field, title field and all fields combined

Figure 2



Key to MARC fields
110 Corporate author
245 Title
260 Publisher
440 Series title
490 Series title
500 General note
505 Contents note
550 Issuing bodies note
650 Subject
710 Added entry, corporate name
740 Added title entry

The chart includes only MARC fields that contained at least one significant term.

Success rate of keyword searches on selected fields from the MARC records